Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Static thresholds and limitations of linear assumptions in reservoir-induced seismicity assessment: Toward a dynamic machine learning framework

Yiming Zhu

153763977616163.com

IGPCEA: Institute of Geophysics China Earthquake Administration

Fengling Yin

IGPCEA: Institute of Geophysics China Earthquake Administration

Changsheng Jiang

IGPCEA: Institute of Geophysics China Earthquake Administration

Jiashan Zhang

IGPCEA: Institute of Geophysics China Earthquake Administration

---

Research Article

Keywords: Reservoir-induced seismicity, Machine learning, Dynamic hazard assessment, Feature importance, Global case database

Posted Date: June 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9686148/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

The applicability of static thresholds and the limitations of linear assumptions in current standards for assessing Reservoir-Induced Seismicity (RIS) hazards require urgent validation. Using one of the most comprehensive global RIS case databases to date (273 cases from 47 countries), this study reveals the limited explanatory power of traditional methods. Pearson and Spearman analyses show weak linear correlations between hydraulic parameters (e.g., dam height, reservoir capacity) or spatiotemporal parameters (e.g., epicentral distance from dams, time lag between impoundment and maximum-magnitude events) and the maximum RIS magnitude (Mmax). Random Forest (RF) feature importance and SHAP correlation analyses indicate that spatiotemporal parameters are much stronger predictors of Mmax than hydraulic factors. A residual-stacking hybrid model (MIXED) combining RF and Gradient Boosting Decision Trees (GBDT) achieves modest but optimal predictive performance among tested models, clearly outperforming Support Vector Machines (SVM) and Multiple Linear Regression (MLR). However, extreme events remain systematically underestimated because of missing physical parameters and limited high-magnitude samples (only 9 of 273 cases have Mmax ≥ 6.0, and only 2 were available for modeling). We further propose a conceptual framework for dynamic hazard assessment based on machine learning-derived probabilities, spatiotemporal parameters reflecting cumulative pore-pressure diffusion, and iterative threshold optimization through database updates.

1. Introduction

Reservoir-Induced Seismicity (RIS) threatens regions surrounding large water projects. Strong reservoir-induced earthquakes can result in catastrophic consequences. According to USCLD (1997), four global cases exceed M6.0: the 1962 Xinfengjiang (M6.0), 1963 Kariba (M6.25), 1966 Kremasta (M6.3), and 1967 Koyna (M6.3) earthquakes, with the latter causing over 200 fatalities (Gupta, 2018). Global construction of large reservoirs has accelerated since 1985. Human activities have increased global lake water storage by ~ 4% and surface area by ~ 11% (Messager et al., 2016), with approximately 4.4 million reservoirs worldwide holding over 8 trillion cubic meters of water (Lehner et al., 2024). In China, the lower Jinsha River cascade development (2015–2021) added four hydropower stations with combined storage exceeding 50 billion cubic meters. Reported RIS cases have increased globally, particularly in Brazil, India, and Canada, likely due to enhanced detection from expanded monitoring networks and regional seismogenic variations (Lamontagne et al., 2018). Comprehensive seismic hazard assessments are imperative.

Current RIS hazard assessments predominantly rely on factor-based and probabilistic evaluation methods. For instance, China's National Standard GB 21075 – 2007, Reservoir-Induced Seismic Hazard Assessment, released in 2008, outlines ten earthquake-inducing factors and employs conditional probability to estimate the likelihood of induced earthquakes. This reliance may be attributed to two key factors. First, despite rapid advancements in physics-based models and numerical simulation techniques—such as RIS hazard assessment methods utilizing poroelastic scenarios and three-dimensional finite element models (Sun et al., 2014), as well as fully coupled pore-viscoelastic equations

(Wang et al., 2024)—these approaches are still limited in their ability to precisely model and characterize the fluid infiltration mechanisms induced by reservoir impoundment and discharge operations and its interaction with faults. Additionally, their generalizability and practical implementation face significant limitations. Second, RIS hazard analysis continues to confront multifaceted and systemic challenges. For example, the precise detection and identification of seismogenic faults remain highly challenging, particularly for faults associated with *M*4.0 earthquakes capable of significantly impacting reservoir-adjacent areas. These faults, often only kilometers in length and tens of meters in width, frequently exceed the resolution limits of conventional reflection seismic surveys (Nantanoi et al., 2022). Furthermore, determining whether pre-existing faults are in a critical state poses another challenge, as even minuscule perturbations can induce seismicity (McGarr et al., 2002).

However, such methods face numerous challenges. Firstly, they heavily rely on linear assumptions and static thresholds (Barros et al., 2018). For example, China's GB 21075 − 2007 uses dam height ≥ 100 m and capacity ≥ 5×10^8 m^3 as thresholds, assuming parameter independence and linear dependency with M_{max}. Similarly, probabilistic methods also rely on linear relationships. Secondly, these methods face challenges related to complex physical mechanisms. RIS involves intricate interactions among shear stress, normal stress, and pore fluid pressure on fault surfaces (Gahalaut et al., 2016). Factors influencing reservoir-induced seismicity may include the ambient stress field conditions in the reservoir area, the characteristics of pre-existing faults, the mechanical and hydraulic properties of rocks (such as strength, frictional resistance, porosity, and permeability), regional geological conditions, and the characteristics of water level fluctuations (Talwani, 1997). However, information related to these physical mechanisms is difficult to obtain and integrate into practical engineering methodologies. Thirdly, these methods are challenged by the sufficiency of information regarding reservoir-induced earthquake cases and earthquake-inducing factors. Currently, the number of cases available for extracting regular patterns is relatively limited. For instance, the database released by the U.S. Geological Survey (USGS) in 1983 contains only 66 recognized cases of reservoir-induced earthquakes (Perman et al., 1983), and the number of cases mentioned in Gupta’s 1992 book is similar. The U.S. Bureau of Reclamation (USBR) listed 40 frequently cited cases of reservoir-induced earthquakes in its 1997 report on reservoir-induced seismicity (USCLD, 1997). During the compilation of China’s national standard GB 21075 − 2007, only about 120 cases were used.

Throughout this manuscript, we use the term reservoir-induced seismicity (RIS) as a general descriptor for seismic events associated with reservoir impoundment, encompassing both reservoir-induced (nucleation of new earthquakes) and reservoir-triggered (advancing the timing of naturally occurring earthquakes) mechanisms, unless specifically noted otherwise. The distinction between these terms involves complex physical debates beyond the scope of this statistically-oriented study. In response to the aforementioned challenges in reservoir-induced seismicity hazard assessment methods, this study aims to explore potential improvements through the following research approaches. Firstly, we will comprehensively collect and organize publicly reported global cases of reservoir-induced earthquakes, along with available information on earthquake-inducing factors, to systematically quantify the

correlations, independence, and significance of hydraulic and spatiotemporal parameters in relation to the maximum magnitude (M_{max}). Secondly, we will develop hybrid prediction models based on machine learning methods and ensemble learning frameworks to explore their potential in overcoming the performance limitations of traditional linear approaches. Finally, we will investigate new frameworks and methodologies for hazard assessment. The exploration of this research pathway will provide new tools for water conservancy engineering hazard management that integrate physical rationality and engineering feasibility.

2. Collection and organization of reservoir event cases

To systematically investigate the analytical framework for RIS hazard assessment and the relationships between maximum magnitude (M_{max}) and key factors such as reservoir capacity, average water depth, and surface area, this study compiled 273 RIS cases from publicly reported literature. These cases encompass diverse reservoir scales, geological settings, and seismic characteristics, providing comprehensive data support for understanding inducing mechanisms and factors influencing M_{max}. Each case was documented with multidimensional parameters, including reservoir name, country, dam height, reservoir capacity, surface area, epicentral distance from the dam, initial impoundment date, timing of M_{max} occurrence, maximum magnitude (M_{max}), and lithology of the epicentral region. Data sources included peer-reviewed journal articles, monographs, and technical reports, all rigorously vetted for scientific credibility. The compiled cases were standardized into a structured table (see Table S1 in Supplementary Materials) to facilitate statistical analysis. Cross-referencing of literature ensured parameter consistency and accuracy.

The dataset spans 47 countries, with reservoir capacities ranging from 0.001×10^8 m^3 to 1689×10^8 m^3, surface areas from 0.285 km^2 to 8,482 km^2, and M_{max} values from microseismic events to M_w7.1, reflecting the diversity of reservoir scales and geological environments. Among the cases, dam heights of 149 reservoirs are exceeding 100 m, with 32 surpassing 200 m, highlighting the prevalence of large dams. Epicentral lithologies include granite, gneiss, limestone and sandstone, underscoring the geological heterogeneity of RIS-prone areas. RIS events predominantly occur in regions with fractured rock masses, active fault systems, or high pore-pressure diffusion. For example, gneiss-dominated areas (*e.g.*, Castanhão Reservoir, Brazil) frequently exhibit low-magnitude seismicity, emphasizing the role of rock physical properties. Epicentral distances from dams vary widely, from 0.1 km to tens of kilometers. For instance, the M_{max} event at Koyna Reservoir (India) occurred 2 ~ 3 km downstream of the dam, while other large reservoirs recorded M_{max} events tens of kilometers away. These spatial variations likely reflect interactions among lithology, tectonic stress fields, and pore-pressure diffusion pathways. Collectively, the data provide a critical foundation for unraveling the complex interplay between reservoir characteristics and RIS maximum magnitude.

Due to the complexity of reservoir shapes, there exist various forms such as gently sloping basin-type reservoirs and V-shaped canyon-type reservoirs located in high mountain and canyon regions. Even

when the water surface area and reservoir capacity are similar, the elastic strain or hydrostatic pressure generated at the bottom of differently shaped water storage areas can vary significantly. To address this, we have defined the parameter of average water depth by the ratio of reservoir capacity to water surface area, enabling a more quantitative statistical correlation analysis. Taking both the Huangshi Reservoir and the Hunanzhen Reservoir in China as examples, their water surface areas are 43.165 km^2 and 46.759 km^2, but their average water depths are 14.18 m and 44.06 m, respectively. The corresponding dam height differences are even more pronounced, at 40.5 m and 129 m, indicating that the average water depth and dam height together reflect different physical characteristics of the reservoir area. Therefore, we introduce the average water depth as a new potential earthquake-inducing factor for analyzing the hazard of reservoir-induced seismicity.

Through the systematic data collection and organization of this study, we have established one of the most comprehensive RIS case databases globally. This database not only provides richer data support for in-depth analysis of the mechanisms by which factors such as reservoir capacity and water surface area affect M_{max} but also offers important theoretical support and practical reference for RIS hazard assessment.

3. Analysis of correlation and importance of earthquake-inducing factors

3.1. Correlation analysis between earthquake-inducing factors and maximum magnitude

The prediction of RIS magnitude requires a systematic evaluation of the statistical relationships between various engineering parameters and seismic responses. This study employs a combination of parametric tests (Pearson correlation coefficient) and non-parametric tests (Spearman rank correlation) to comprehensively analyze the correlations between six key parameters and the maximum magnitude (M_{max}) in a semi-logarithmic coordinate system. The parameters include: Dam Height, Reservoir Capacity, Reservoir Area, Average Water Depth, Distance from Epicenter to Dam, Time Lag between Maximum Magnitude Event and First Impoundment.

Variable definitions: Time lag (Δt) is defined as the number of days from the initial reservoir impoundment date to the date of the M_{max} event. Epicentral distance (D) is defined as the great-circle distance from the dam coordinate to the epicenter of the M_{max} event, measured in kilometers. Both parameters were log-transformed (base 10) for correlation analysis to accommodate their wide ranges and right-skewed distributions, consistent with the treatment of hydraulic parameters. Pearson correlation coefficient is a statistical measure based on linear relationships, calculated as the standardized covariance to quantify the strength of linear correlation. Its value ranges from [-1, 1], where positive values indicate a positive correlation, and negative values indicate a negative correlation (Pearson, 1895). Its mathematical expression is:

$$r_p = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

1

Spearman correlation coefficient is a non-parametric measure based on rank order, designed to capture monotonic (nonlinear) relationships. It is particularly suitable for non-normally distributed data or datasets containing outliers (Spearman, 1904). Its mathematical expression is:

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

2

In the above equation, *x* represents the logarithmically transformed earthquake-inducing factor, *y* denotes the maximum magnitude, and d_i signifies the rank difference. The combined use of these two methods enables the identification of linear trends while capturing nonlinear monotonic relationships, complemented by significance testing ($p < 0.05$ for significance) to ensure robust conclusions.

Figure 2 presents scatter plots, trend analysis, and correlation analysis for six earthquake-inducing factors and the maximum magnitude of induced earthquakes. The results indicate that the correlations between these factors and the maximum magnitude are generally weak. The absolute values of the Pearson correlation coefficient r_p range from 0.06 to 0.28, with poor significance levels. Only the time lag (days) between the maximum magnitude event and the initial reservoir impoundment meets the strict criteria for statistical significance, although r_p is merely 0.28. Spearman analysis reveals a similar trend, demonstrating the methodological robustness of the two correlation coefficient evaluations employed in this study. These findings suggest that there is no simple linear relationship between individual earthquake-inducing factors and the maximum magnitude of induced earthquakes, indicating that individual factors have limited standalone explanatory power and limiting their practical predictive utility for threshold-based assessment. It is necessary to consider the complex nonlinear relationships and multifactorial coupling effects among these factors. Furthermore, the correlation between the maximum magnitude event and the time lag of initial reservoir impoundment highlights the long-term effects of reservoir water infiltration and fault activation in the surrounding areas. This challenges the traditional focus on the rapid response phase immediately after initial impoundment in seismic hazard assessment, as well as the static hazard analysis limited to a narrow range around the reservoir (*e.g.*, within 10 km of the reservoir shoreline).

3.2. Importance and independence analysis of earthquake-inducing factors

To reveal the importance of each earthquake-inducing factor in predicting the maximum magnitude, as well as their independent contributions and interaction effects, this study considers both the global influence and local interpretability of the features. Two complementary metrics—Random Forest (RF) built-in feature importance and SHAP (SHapley Additive exPlanations) values—were employed for analysis (Fig. 3a, b). The combination of these two approaches provides insights into the different mechanisms of feature contributions in the model. Specifically, RF importance reflects the extent to which the absence of a feature degrades model performance, while SHAP values characterize the specific direction and magnitude of a feature's influence on predictions.

Important note on interpretation: The feature importance metrics employed in this study (RF permutation importance and SHAP values) identify statistical predictive relevance, not physical causality. The models do not explicitly incorporate hydraulic diffusivity, permeability, fault geometry, in-situ stress state, or reservoir operation history. Therefore, the relative importance of time lag and epicentral distance should be understood as reflecting their statistical information content for M_{max} prediction within the available parameter space, rather than as direct evidence for specific physical mechanisms such as pore-pressure diffusion rates or fault activation processes. Causal inferences would require coupled hydro-mechanical modeling beyond the scope of this statistically-oriented framework.

First, the Random Forest (RF) algorithm, based on an ensemble of decision trees, was used to rank feature importance. The RF importance (OOB Permutation Importance) evaluates the global influence of features by measuring the change in prediction error before and after feature permutation. This method constructs multiple decorrelated decision trees and uses out-of-bag (OOB) data to calculate the feature permutation importance metric. Specifically, the relative increase in model prediction error (mean squared error) is observed after randomly permuting a feature's values. The mathematical expression for this is:

$$Importance\left(f_i\right)=\frac{1}{N_{tree}}\sum_{t=1}^{N_{tree}}\left(MSE_{perm}^{(t)}\left(f_i\right)-MSE_{oob}^{(t)}\right)$$

3

where $MSE_{oob}^{(t)}$ represents the out-of-bag (OOB) error of the t-th tree, and $MSE_{perm}^{(t)}\left(f_i\right)$ denotes the error after permuting the values of feature f_i. This metric eliminates the influence of feature scales and effectively identifies features with predictive value in nonlinear relationships. In the specific calculation process, we set the Monte Carlo sampling number $n_{samples}$=200 to balance computational efficiency and accuracy. The hyperparameters of the random forest were determined through Bayesian optimization (number of trees $N_{trees}\in$[50, 500], minimum samples per leaf node $n_{leaf}\in$[1, 50], with the optimization objective being the minimization of the 5-fold cross-validated *RMSE*).

Next, we used SHAP values to describe feature importance. SHAP values are based on cooperative game theory and quantify the independent influence of a feature by calculating the average marginal contribution of the feature across all possible subsets. Their advantage lies in satisfying the additivity

axiom, enabling the separation of interaction effects between features. The mathematical expression for SHAP value calculation is:

$$\varnothing_j = \frac{1}{N}\sum_{i=1}^{N}\left|\varnothing_j^{(i)}\right|$$

4

We further computed the SHAP interaction matrix to analyze the independence among the six earthquake-inducing factors. The SHAP interaction matrix is based on the joint effects of feature values on the prediction outcomes. By pairwise computation of the six earthquake-inducing factors and representing the standardized correlation coefficients ρ_{jk} through color mapping:

$$\rho_{jk} = \frac{\mathrm{E}\left[\left(\phi_j - \bar{\phi}_j\right)\left(\phi_k - \bar{\phi}_k\right)\right]}{\sigma_{\phi_j}\sigma_{\phi_k}}$$

5

Figure 3a, b present the SHAP importance and RF importance results for the six earthquake-inducing factors, respectively. The SHAP importance values for the six factors range from 0.33 to 0.36, showing minimal variation among them. In contrast, the RF importance values span from 0.21 to 0.69, indicating significant differences among the six factors. Ranked from highest to lowest, these factors are: distance between the epicenter of the maximum magnitude event and the dam (0.69), time delay between the maximum magnitude event and initial reservoir impoundment (0.58), average reservoir water depth (0.34), dam height (0.30), reservoir capacity (0.27), and reservoir area (0.21). The notable discrepancy between the two importance evaluation methods can be attributed to the divergent model dependency paths induced by the six factors. In the RF importance calculation, the distance between the epicenter of the maximum magnitude event and the dam causes a sharp increase in prediction error, indicating that this feature is a critical variable for model splitting nodes. On the other hand, the balanced distribution of SHAP values for the six factors suggests that the model relies more on the synergistic effects of multiple features during individual predictions, aligning with the multi-factor coupling mechanism of reservoir-induced seismicity. Notably, the importance of traditional hydraulic parameters (dam height, reservoir capacity, reservoir area, and average reservoir water depth) is lower than that of spatiotemporal parameters (time delay between the maximum magnitude event and initial reservoir impoundment, and distance between the epicenter of the maximum magnitude event and the dam). This statistical finding indicates that, within the available parameter space, spatiotemporal parameters carry more predictive information for M_{max} than hydraulic parameters, though this does not imply that hydraulic factors are physically unimportant, rather, their effects may be mediated through unobserved variables such as fault properties and stress states

The SHAP value correlation matrix for independence analysis of the six earthquake-inducing factors is shown in Fig. 3c. The correlation matrix reveals that the SHAP correlation coefficients for reservoir

capacity, dam height, average reservoir water depth, and reservoir area are as high as 0.97 ~ 0.99, indicating that these parameters share nearly identical sources of information and effectiveness in the model. This suggests a high degree of correlation and statistical coupling among these four hydraulic-related factors. The redundancy of these parameters implies that easily obtainable proxy variables (*e.g.*, using dam height as a substitute for water depth) can be prioritized in practical monitoring. In contrast, the two spatiotemporal parameters—distance between the epicenter of the maximum magnitude event and the dam, and time delay between the maximum magnitude event and initial reservoir impoundment—exhibit stronger independence. The SHAP correlation coefficient between these two parameters is approximately 0.90, and their SHAP correlation coefficients with the other four hydraulic-related factors range from 0.94 to 0.96, indicating that their influence mechanisms are statistically distinguishable from hydraulic parameters. This supports a separate modeling strategy for statistically independent factors, such as the temporal effects of fluid infiltration and the spatial distribution of activated faults.

4. Maximum magnitude prediction using an ensemble learning framework

4.1. Prediction models and ensemble learning framework

Based on the previous analysis of the correlation, independence, and importance of earthquake-inducing factors with respect to M_{max}, it is evident that relying solely on individual earthquake-inducing factors or linear relationships between these factors and M_{max} is insufficient for effective prediction. Therefore, we further explore the potential of a hybrid model based on an ensemble learning framework to extract complex nonlinear relationships and improve the prediction of the maximum magnitude (M_{max}) of reservoir-induced seismicity.

The technical approach of this study, which employs machine learning techniques to predict M_{max}, is designed based on the following considerations: The nonlinear characteristics of reservoir-induced seismic systems require algorithms with strong nonlinear modeling capabilities. The limited sample size (n = 120) necessitates models with anti-overfitting properties. Engineering applications demand models that balance prediction accuracy with physical interpretability. To address these requirements, this study utilizes Random Forest (RF) and Gradient Boosting Decision Tree (GBDT) as base models. The Bayesian optimization algorithm is employed to determine the optimal hyperparameter combinations for RF and GBDT, and residual stacking is used to construct a new hybrid model. For comparative purposes, we selected Multiple Linear Regression (MLR) and the classic machine learning model Support Vector Machine (SVM) as baseline models, enabling comparisons from both linear and nonlinear perspectives. All models undergo automatic hyperparameter tuning via Bayesian optimization and are evaluated using a unified 5-fold cross-validation framework to ensure comparability.

Random Forest (RF) is chosen as the primary base model due to its significant advantages. By employing Bootstrap aggregation and the random subspace method, RF effectively reduces model

variance, which is crucial for small sample conditions (Breiman, 2001). RF utilizes the Bagging method of ensemble learning and the random subspace technique, which notably lowers variance, making it particularly suitable for small datasets. Its built-in Out-of-Bag (OOB) error evaluation provides a more stable basis for parameter optimization. Additionally, decision trees are non-parametric models that do not require predefined data distributions, allowing them to better adapt to the non-stationary characteristics of RIS systems, which exhibit varying inducing mechanisms across different geological contexts. Furthermore, RF can quantify feature importance through the OOB Permutation Importance method, thereby identifying key earthquake-inducing factors. The hyperparameter optimization of the RF model is achieved through a Bayesian search framework. Three key parameters are defined: the number of decision trees (N_{trees}, range 50 ~ 500), the minimum number of samples per leaf node (*MinLeaf*, range 1 ~ 50), and the number of features randomly sampled for each tree (N_{features}, range 1 ~ 6). The OOB error is used as the optimization objective function. The Bayesian optimization process executes 30 iterations, with 5 parallel candidate parameter combinations generated per iteration, ultimately selecting the parameter configuration that minimizes the OOB-MSE.

The introduction of Gradient Boosting Decision Trees (GBDT) is based on their residual optimization characteristics. The gradient boosting framework proposed by Friedman (2001) iteratively corrects prediction errors, making it particularly suitable for extreme value prediction tasks. The GBDT model adopts a layer-by-layer residual fitting strategy, with 20 rounds of optimization to determine the number of decision trees, learning rate, and maximum splitting depth. The hyperparameter optimization for GBDT focuses on three core parameters: the number of weak learners (N_{trees}, range 50 ~ 500), the learning rate (η, log-transformed range 0.01 ~ 0.3), and the maximum number of splits per tree (*MaxSplits*, range 5 ~ 50). A nested cross-validation strategy is employed to evaluate parameter performance, where the inner 5-fold cross-validation calculates the Root Mean Square Error (*RMSE*), and the outer optimizer performs 20 iterations of search based on a Gaussian process surrogate model. The final model adopts the optimal parameter combination that minimizes the *RMSE* on the validation set.

Based on the RF and GBDT models and their optimization, we propose a Residual Stacking Hybrid Model (hereinafter referred to as the Mixed Model). Since the RF model employs an implicit global interaction approach for feature interactions, it reduces variance through Bagging and is suitable for coarse-grained pattern recognition. However, it is susceptible to sampling bias when predicting extreme values (*e.g.*, maximum magnitude). On the other hand, GBDT utilizes explicit local interactions and fine-grained residual correction, reducing bias through Boosting and optimizing extreme value prediction through continuous residual correction. The combination of RF and GBDT enables joint bias-variance optimization and enhances feature interactions, and their complementarity forms the theoretical foundation of the Mixed Model. We adopt a two-stage residual stacking architecture: first, the RF model generates initial predictions $\hat{y}_{RF}$, and the residuals between the true values and the predictions $residual = y_{\text{ture}} - \hat{y}_{RF}$ are calculated. Subsequently, the RF predictions are used as additional features input into the GBDT model to fit the residuals, yielding the final predictions of the Mixed Model $\hat{y}_{Hybvrid} = \hat{y}_{RF} + \hat{y}_{Residual}$. The hyperparameters of the residual model, including the number of trees

(50 ~ 300), learning rate (0.05 ~ 0.2), and maximum number of splits (10 ~ 30), are independently determined through Bayesian optimization.

As the first baseline model for comparison, the Multiple Linear Regression (MLR) model is based on the assumption of linear relationships between variables, constructing a linear equation to describe the relationship between the dependent variable (M_{max}) and multiple independent variables (earthquake-inducing factors). Its mathematical expression is $\hat{y}_{MLR} = \beta_0 + \sum_{i=1}^{6} \beta_i x_i$, where β_0 is the intercept, and β_i are the regression coefficients corresponding to the independent variables, measuring the impact of each independent variable on the dependent variable. The MLR model is characterized by its simplicity, intuitive principles, high computational efficiency, and ease of interpretation. The second baseline model for comparison is the Support Vector Machine (SVM). SVM is a machine learning method based on structural hazard minimization theory, particularly suitable for nonlinear mapping and regression analysis in small-sample, high-dimensional spaces (Vapnik, 1995; Das et al., 2012; Asim et al., 2018). We selected the RBF kernel function and optimized the kernel width (KernelScale) and regularization parameter (C) through grid search. The linear inseparability of SVM makes it a benchmark for nonlinear modeling. In this study, the optimized parameters for the SVM model include the penalty coefficient C ($1e^{-3}$~$1e^{3}$), kernel bandwidth σ ($1e^{-3}$~$1e^{3}$), and insensitive band ε (0.01~1), which are determined through 30 rounds of Bayesian optimization, with generalization error calculated using 5-fold cross-validation.

4.2. Comprehensive evaluation framework for prediction models

To comprehensively evaluate the performance of the reservoir-induced earthquake maximum magnitude prediction models, we first employed four statistical metrics to assess the prediction results. These include: Pearson Correlation Coefficient: This metric reflects the strength of the linear relationship between predicted and observed values. Root Mean Square Error (*RMSE*): *RMSE* characterizes the dispersion of prediction errors. Due to the squaring operation, it amplifies the impact of larger errors, thus placing greater emphasis on penalizing extreme prediction deviations. Its units are consistent with the original magnitude data (Richter scale), making it directly interpretable for engineering hazard assessment. Mean Absolute Error (*MAE*): *MAE* calculates the average of the absolute values of prediction errors, assigning linear weights to all errors. Compared to *RMSE*, it is more robust to outliers and provides a more stable reflection of the prediction accuracy for typical samples. Coefficient of Determination (R^2): R^2 quantifies the model's ability to explain the variance in magnitude data. Its value ranges from [0, 1], representing the proportion of error reduced by the model compared to simply using the mean for prediction. For example, $R^2 = 0.35$ indicates that the model explains 35% of the magnitude variance, making it a core metric for assessing the goodness of fit. It is worth noting that, considering the small sample size and the long-tailed distribution characteristics of earthquake magnitude data, this study applied robust standardization to the evaluation metrics. Specifically, the median and interquartile range (IQR) were used instead of the mean and standard deviation for data normalization, thereby

mitigating the interference of extreme values on model training. Subsequently, the model outputs were inverse-transformed to restore them to the original magnitude scale.

In addition to statistical metrics, this study utilized visualization techniques to further analyze the performance differences among the five prediction models. We plotted the probability density curves of prediction errors for each model using Kernel Density Estimation (KDE), which fits the distribution of prediction errors. The mathematical expression for KDE is:

$$\hat{f}_h = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{e - e_i}{h}\right)$$

6

Here, *h* represents the kernel bandwidth, and *K*($\frac{e-e_i}{h}$) is the Epanechnikov kernel function. In the error probability density plot, the peak position reflects the central tendency of the errors (a leftward shift indicates systematic underestimation), while the width of the curve characterizes the dispersion of the errors (a narrow band suggests high precision). Additionally, we used boxplots of the statistical metrics to visualize their statistical characteristics. Specifically, the interquartile range (IQR, 25%–75%) was used to measure the variability of the metrics, and the median line was used to represent the overall performance of the models. A lower median indicates better performance for *RMSE* and *MAE*, while a higher median indicates better performance for the coefficient of determination (R^2). A smaller IQR suggests more stable prediction results, and fewer outliers indicate stronger robustness of the model.

4.3. Results analysis

The 5-fold cross-validation results of the hybrid model (MIXED) based on the ensemble learning framework are shown in Fig. 4. The figure also presents the independent prediction results of the RF and GBDT models, which constitute the hybrid model. For comparison, the prediction results of the two baseline models, SVM and MLR, are also provided.

The Pearson correlation coefficient (r_p) was used to evaluate the correlation between the predicted and observed magnitudes. As shown in Fig. 4a ~ e, the hybrid model (MIXED) demonstrates the closest alignment between predicted and observed magnitudes, with the highest Pearson correlation coefficient (r_p =0.494). The RF and GBDT models follow with r_p =0.486 and r_p =0.438, respectively, while the SVM and MLR models exhibit lower correlations, with r_p values of 0.062 and 0.135, respectively. From the probability density plots of prediction errors (Fig. 4f ~ j), the error distributions of the MIXED, RF, and GBDT models are similar, with peaks close to the standardized zero line. In contrast, the SVM and MLR models show higher error dispersion, and their peaks deviate significantly from the zero line.

The boxplot of *RMSE* (Fig. 4k) reveals that the MIXED model has the lowest median *RMSE* (0.998), indicating the smallest dispersion of prediction errors. The other models, ranked from low to high error dispersion, are the RF model (*RMSE* = 1.035), GBDT model (*RMSE* = 1.116), MLR model (*RMSE* = 1.190),

and SVM model (*RMSE* = 1.207). Similarly, the boxplot of *MAE* (Fig. 4l) shows that the MIXED model has the lowest median *MAE* (0.808), indicating the smallest average absolute prediction error. The other models, ranked from low to high average absolute error, are the RF model (*MAE* = 0.845), GBDT model (*MAE* = 0.917), MLR model (*MAE* = 0.976), and SVM model (*MAE* = 1.073). The boxplot of R^2 (Fig. 4m) indicates that the MIXED model has the highest median (R^2 = 0.306), representing the greatest reduction in error. The other models, ranked by error reduction, are the RF model (R^2 = 0.244), GBDT model (R^2 = 0.160), while the MLR and SVM models exhibit negative R^2 values.

The evaluation results clearly demonstrate that the hybrid model (MIXED) significantly outperforms the baseline models (SVM and MLR) across all four core metrics and slightly outperforms the RF and GBDT models within the ensemble learning framework. This indicates that the hybrid model based on the ensemble learning framework can significantly improve the prediction accuracy of M_{max}, showcasing its high potential in reservoir-induced seismic hazard analysis. Meanwhile, the failure of the MLR model as a baseline highlights the inability of linear assumptions to capture the complex nonlinear relationships between earthquake-inducing factors and M_{max}.

5. Discussion

5.1. Suggestions for improving the classification of reservoir-induced seismic hazard assessment

China's GB 21075 − 2007 standard defines RISHA assessment areas as reservoir regions plus 10 km periphery, using dam height (≥ 100 m) and capacity (≥ 5×10^8 m^3) thresholds, with two classification levels (Class A/B). The standard identifies 10 earthquake-inducing factors through fault investigations and deterministic M_{max} evaluations, then integrates probabilistic assessments. However, based on only 120 cases, threshold rationality must be revisited. This study uses 273 cases to analyze six factors' correlations and independence, evaluating five machine learning models for M_{max} prediction to explore improved thresholds.

The following shortcomings are identified: (1) Linear independence failure: SHAP value correlation analysis (Fig. 3c) shows correlation coefficients of 0.97 ~ 0.99 among dam height, capacity, area, and depth, indicating strong statistical coupling, yet GB 21075 − 2007 treats height and capacity as independent, causing redundant calculations. (2) Missing spatiotemporal parameters: RF analysis (Fig. 3b) shows time lag (importance: 0.58) and epicenter-dam distance (0.69) outperform traditional parameters (height: 0.30, capacity: 0.27), but GB 21075 − 2007 lacks temporal features, ignoring fluid infiltration's cumulative effects. (3) Nonlinear effects neglected: Fixed thresholds ignore nonlinear interactions, causing over/underestimation. Among 242 cases with complete data, combined thresholds missed 23/103 M_{max}≥4.0 events (22.3% miss rate), 7/33 M_{max}≥5.0 (21.2%), and 2/9 M_{max}≥6.0 (22.2%), necessitating improved classification.

Future improvements require: (1) Conceptual dynamic multi-level framework using ML-predicted probability P(M_{max}≥4.0) as induction condition with formula P = $S(\beta_1 \ln(\text{Dam Height}) + \beta_2 \ln(\text{Reservoir Capacity}) + \beta_3 \ln(\text{Time Lag}) + \beta_4 \ln(\text{Epicenter Distance}))$, where S is Sigmoid and β are ML coefficients, using P_{95} threshold covering 95% of $M \geq 4.0$ cases. This formula is presented as a conceptual prototype for discussion; its coefficients require fitting and validation on expanded datasets before operational implementation. (2) Classification index defined under operation period and impact zone constraints. (3) Dynamic update mechanism introducing time-varying hazard coefficients, updating parameters based on operation period and reassessing every 5 years. (4) Comprehensive database for continuous incremental learning and threshold optimization.

5.2. Engineering implications of the maximum magnitude prediction model based on the ensemble learning framework

The MIXED model has limitations for M_{max} prediction and engineering applications: (1) Insufficient factor coverage: Only six accessible parameters are incorporated, neglecting critical mechanical parameters like fault friction and permeability anisotropy. (2) Statistical-proxy limitation: Time lag is input as static feature without coupling pore pressure diffusion equations, inadequately describing permeability-fault dynamics. Its predictive relevance reflects statistical association with unresolved physical processes rather than direct causal representation. (3) High-magnitude sample scarcity: Only 9/273 cases have M_{max}≥6.0 (with merely 2 cases in the 120-case modeling subset), causing systematic underestimation at high magnitudes (Fig. 4a ~ e). This represents a fundamental constraint: the model cannot reliably predict beyond the observed magnitude distribution. (4) SVM model failure: SVM's global optimization struggles with discrete samples, loses coupling information after dimensionality reduction, is susceptible to outliers, and its Gaussian kernel is less adaptable to multimodal data than decision trees.

M_{max} prediction directly impacts reservoir safety design, hazard assessment, and emergency planning. Though not ideal, the MIXED model (R^2 = 0.306, Fig. 4m) significantly outperforms MLR (R^2 = − 0.12), demonstrating some nonlinear feature extraction capability and providing a preliminary reference for future predictive modeling.

Constructing physically accurate models faces data accessibility challenges, including fault stress states, friction coefficients, permeability, and crustal parameters (Miao et al., 2021). ML methods balance accessibility with physical authenticity. The MIXED model incorporates hydraulic and spatiotemporal parameters (Time Lag, Distance from Dam), which RF analysis shows contribute statistically 58% and 69% to M_{max} prediction, as measured by permutation importance in the available feature space. These values reflect information content for prediction rather than physical effect magnitudes. This justifies ML-based modeling as a pragmatic interim approach for practical reservoir-induced seismicity applications, pending integration of mechanistic parameters..

6. Conclusions

This study addresses the scientific and rational challenges faced by traditional evaluation and hazard analysis methods based on earthquake-inducing factors in the practical engineering application of reservoir-induced seismic hazard assessment. By systematically collecting global reservoir-induced seismic cases and easily accessible earthquake-inducing factors, and employing statistical correlation analysis and machine learning feature importance evaluation, the study analyzes the correlation between earthquake-inducing factors and the maximum magnitude (M_{max}) of reservoir-induced earthquakes, as well as the independence and ranking of these factors, revealing nonlinear coupling relationships. Through ensemble learning modeling and the establishment of a statistical evaluation index system, the feasibility and potential of using machine learning algorithms for M_{max} prediction are explored. The main findings from this study are as follows.

(1) We compiled one of the largest public RIS datasets (273 cases, 47 countries) with four hydraulic parameters and two spatiotemporal parameters. Pearson/Spearman analyses show weak linear correlations with M_{max}. RF and SHAP analyses reveal hydraulic parameter statistical coupling (0.97 ~ 0.99) and relatively independent spatiotemporal mechanisms (0.90 ~ 0.96), suggesting that fluid infiltration diffusion patterns are statistically detectable and critically influence fault activation timing and location.

(2) The MIXED model (RF+GBDT residual stacking) achieves the best performance among tested models (r_p=0.494, *RMSE* = 0.998, *MAE* = 0.808, R^2 = 0.306), significantly outperforming MLR (R^2 = − 0.12). This demonstrates the potential of nonlinear modeling for RIS M_{max} prediction, providing a preliminary engineering tool. However, the moderate absolute performance and high-magnitude sample scarcity (only 2 modeling cases with M_{max}≥6.0) and exclusion of physical parameters like fault friction severely limit extreme event prediction reliability.

(3) Case statistics and model analysis reveal significant practical flaws in current dam height-capacity joint thresholds, with 22.3% miss rate for M_{max}≥4.0 events. Conceptual improvements include: exploring replacement of fixed thresholds with ML-predicted probabilities for dynamic hazard assessment; incorporating spatiotemporal parameters (time lag and epicenter distance) to statistically reflect fluid infiltration's long-term cumulative effects; establishing dynamic database updates for incremental learning and threshold optimization. Traditional conditional probability methods are limited by weak individual factor-M_{max} correlations, making ML methods that balance data accessibility and physical mechanism authenticity a pragmatic interim choice, though not a substitute for physics-based models when sufficient data become available.

Declarations

CRediT Authorship contribution statement

Yiming Zhu: Data curation, Validation, Visualization. **Fengling Yin**: Data curation, Funding acquisition, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Changsheng Jiang**: Conceptualization, Methodology, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Jiashan Zhang**: Data curation, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Special Fund of the Institute of Geophysics, China Earthquake Administration and the program of the China Seismic Experimental Site (grant no. DQJB24Z04). We thank Jiancang Zhuang from the Institute of Statistical Mathematics, for his guidance of data analysis and processing. We thank Peng Han from the Southern University of Science and Technology, for his constructive suggestions on this paper.

Code availability

MATLAB code is available from this GitHub repository: https://github.com/Flyin0726/Code-and-Data.

Data availability

See the Supplemental Material.

References

1. Asim KM, Idris A, Iqbal T, Martínez-Álvarez F (2018) Earthquake prediction model using support vector regressor and hybrid neural networks. PLoS ONE 13(7):e0199004. https://doi.org/10.1371/journal.pone.0199004
2. Assumpção M, Marza V, Barros L et al (2002) Reservoir-induced seismicity in Brazil. The mechanism of induced seismicity. Birkhäuser, Basel, pp 597–617. https://doi.org/10.1007/978-3-0348-8179-1_25
3. Barros LV, Assumpção M, Ribotta LC et al (2018) Reservoir-triggered seismicity in Brazil: Statistical characteristics in a midplate environment. Bull Seismol Soc Am 108(5B):3046–3061. https://doi.org/10.1785/0120170364

4. Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
5. Das SK, Samui P, Sabat AK (2012) Prediction of field hydraulic conductivity of clay liners using an artificial neural network and support vector machine. Int J Geomech 12(5):606–611. https://doi.org/10.1061/(ASCE)GM.1943-5622.0000129
6. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. Ann Stat 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451
7. Gahalaut K, Tuan TA, Rao NP (2016) Rapid and delayed earthquake triggering by the Song Tranh 2 reservoir, Vietnam. Bull Seismol Soc Am 106(5):2389–2394. https://doi.org/10.1785/0120160106
8. Gupta HK (1992) Reservoir-induced earthquakes. Elsevier, Amsterdam [Developments in Geotechnical Engineering, 64]
9. Gupta HK (2018) Review: Reservoir triggered seismicity (RTS) at Koyna, India, over the past 50 yrs. Bull Seismol Soc Am 108(5B):2907–2918. https://doi.org/10.1785/0120180019
10. Lamontagne M, Rogers G, Cassidy J et al (2018) A review of reservoir monitoring and reservoir-triggered seismicity in Canada. Bull Seismol Soc Am 108(5B):3062–3079. https://doi.org/10.1785/0120180009
11. Lehner B, Beames P, Mulligan M et al (2024) The Global Dam Watch database of river barrier and reservoir information for large-scale applications. Sci Data 11:1069. https://doi.org/10.1038/s41597-024-03752-9
12. McGarr A, Simpson D, Seeber L (2002) Case histories of induced and triggered seismicity. Int Handb Earthq Eng Seismol 81:1253–1259
13. Messager ML, Lehner B, Grill G et al (2016) Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nat Commun 7:13603. https://doi.org/10.1038/ncomms13603
14. Miao M, Zhu S, Chang Y et al (2021) Spatiotemporal evolution of pore pressure changes and Coulomb failure stress in a poroelastic medium for different faulting regimes. Earth Space Sci 8. https://doi.org/10.1029/2021EA001837. e2021EA001837
15. Nantanoi S, Rodríguez-Pradilla G, Verdon JP (2022) 3D-Seismic interpretation and fault slip potential analysis from hydraulic fracturing in the Bowland Shale, UK. Pet Geosci 28(2):petgeo2021–petgeo2057
16. Pearson K (1895) Note on regression and inheritance in the case of two parents. Proc R Soc Lond 58:240–242. https://doi.org/10.1098/rspl.1895.0041
17. Perman RC, Packer D, Smith C, Kneupfer PL (1983) Collection of data bank on reservoir-induced seismicity. USGS Contract 14-08-0001-19132
18. Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1):72–101. https://doi.org/10.2307/1412159

19. Sun YJ, Zhang H, Dong SW et al (2014) Study on the effect of Zipingpu reservoir on the occurrence of 2008 Wenchuan MS7.9 earthquake based on a 3D-poroelastic model. Chin J Geophys 57:209–218. https://doi.org/10.1002/cjg2.20097
20. Talwani P (1997) On the nature of reservoir-induced seismicity. Pure Appl Geophys 150:473–492
21. United States Committee on Large Dams (1997) Reservoir triggered seismicity. Denver, CO
22. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York. https://doi.org/10.1007/978-1-4757-2440-0
23. Wang ZT, Zhang H, Guo YC, Meng Q (2024) Numerical simulation of the influence of the Baihetan reservoir impoundment on regional seismicity. Appl Sci 14(12):5145. https://doi.org/10.3390/app14125145

Figures

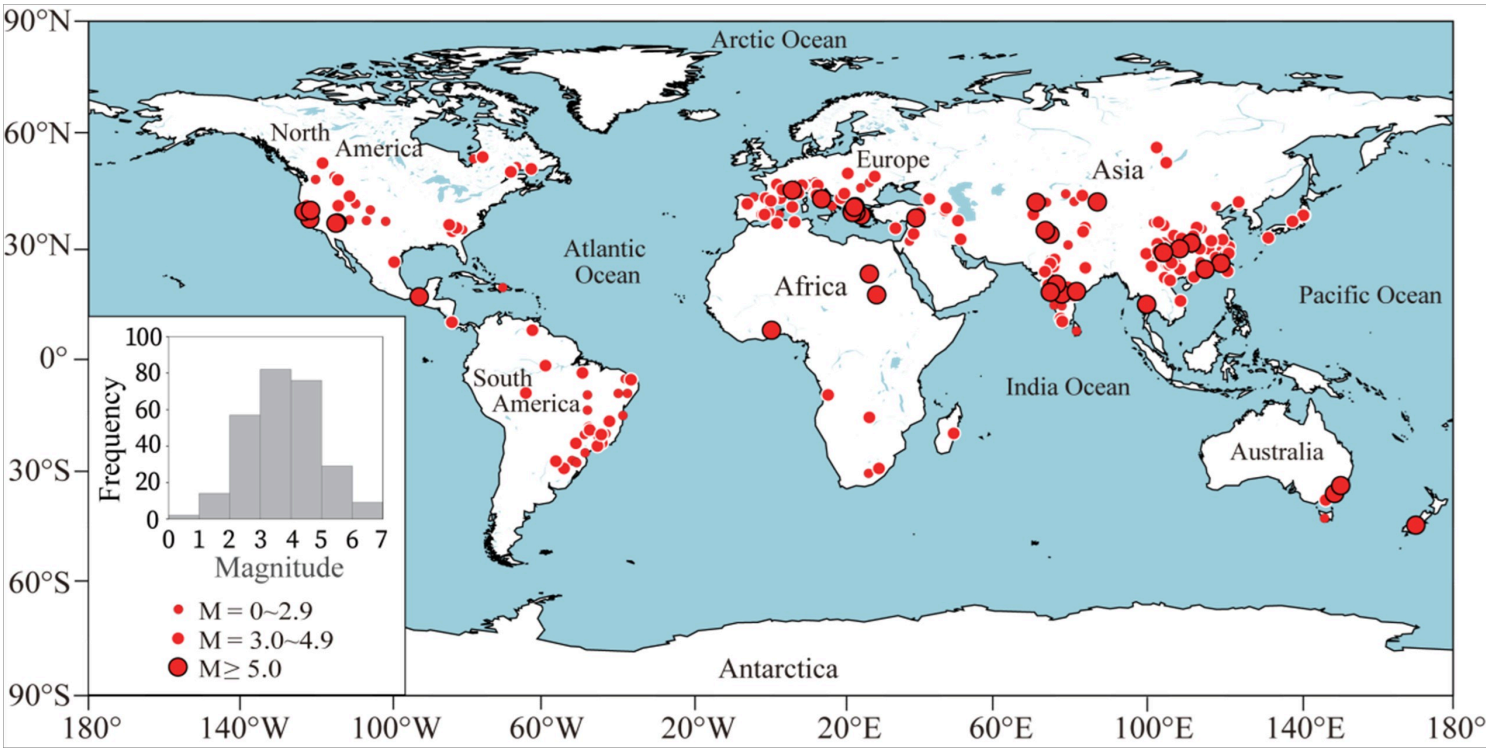


Figure 1

Spatial distribution of reservoir-induced seismicity (RIS) cases collected and organized in this study. The map illustrates the spatial distribution of the maximum magnitude events among the RIS cases, with the inset in the lower left corner showing the frequency distribution of the maximum magnitude (M_{max}).

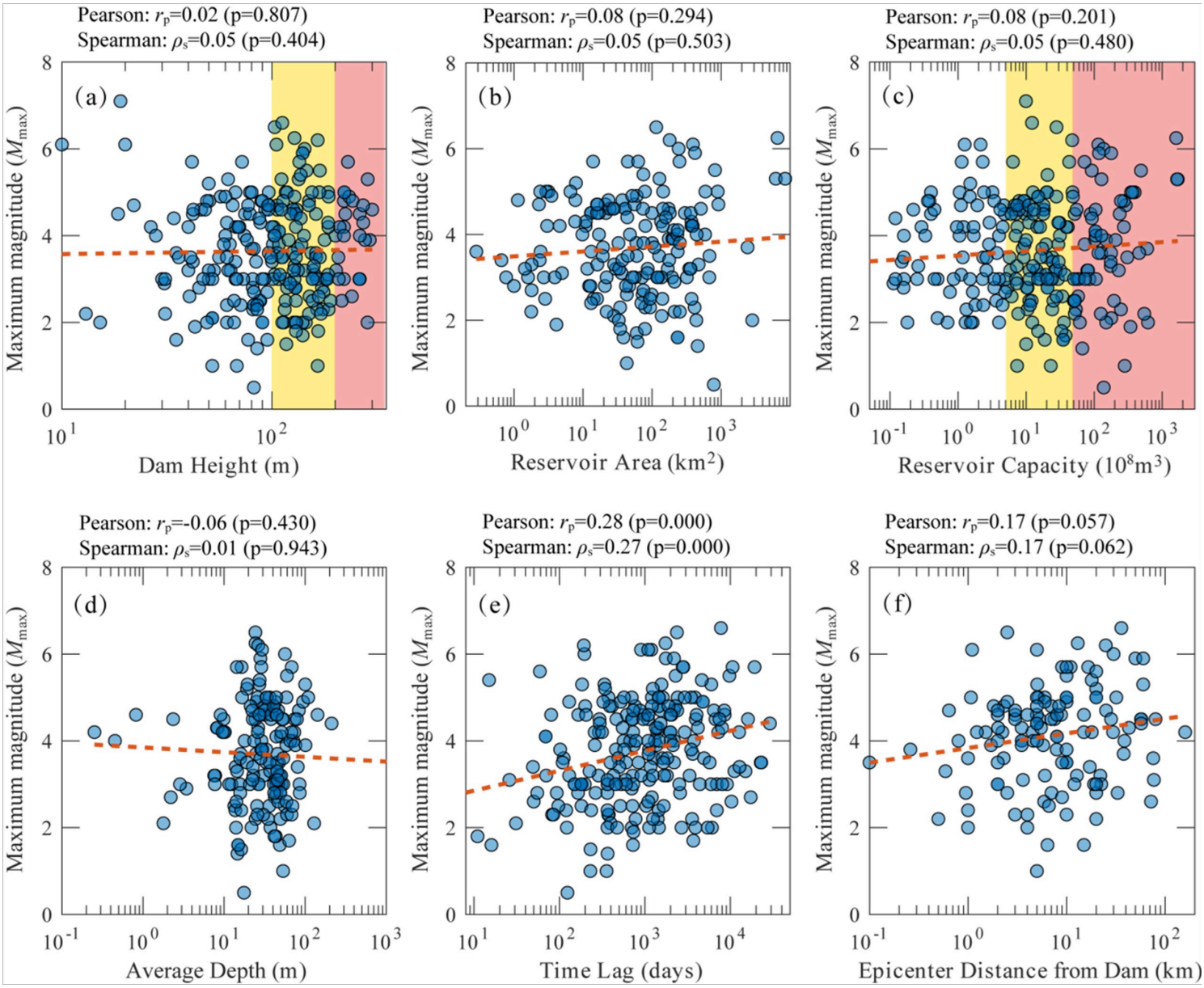


Figure 2

Correlation analysis between earthquake-inducing factors and M_{max}. (a~f) Scatter plots of: (a) dam height, (b) reservoir capacity, (c) reservoir area, (d) average depth, (e) epicenter distance, (f) time lag versus M_{max}. Horizontal axes are logarithmic; vertical axes are linear. Blue circles show observed data (sample sizes in Table S1). Red dashed lines indicate log-transformed linear regression trends. Subplot annotations show Pearson (r_p), Spearman (ρ_s) correlations and p-values. Red and yellow rectangles in (a) and (c) indicate GB/T 21075-2007 Class A and B parameter ranges.

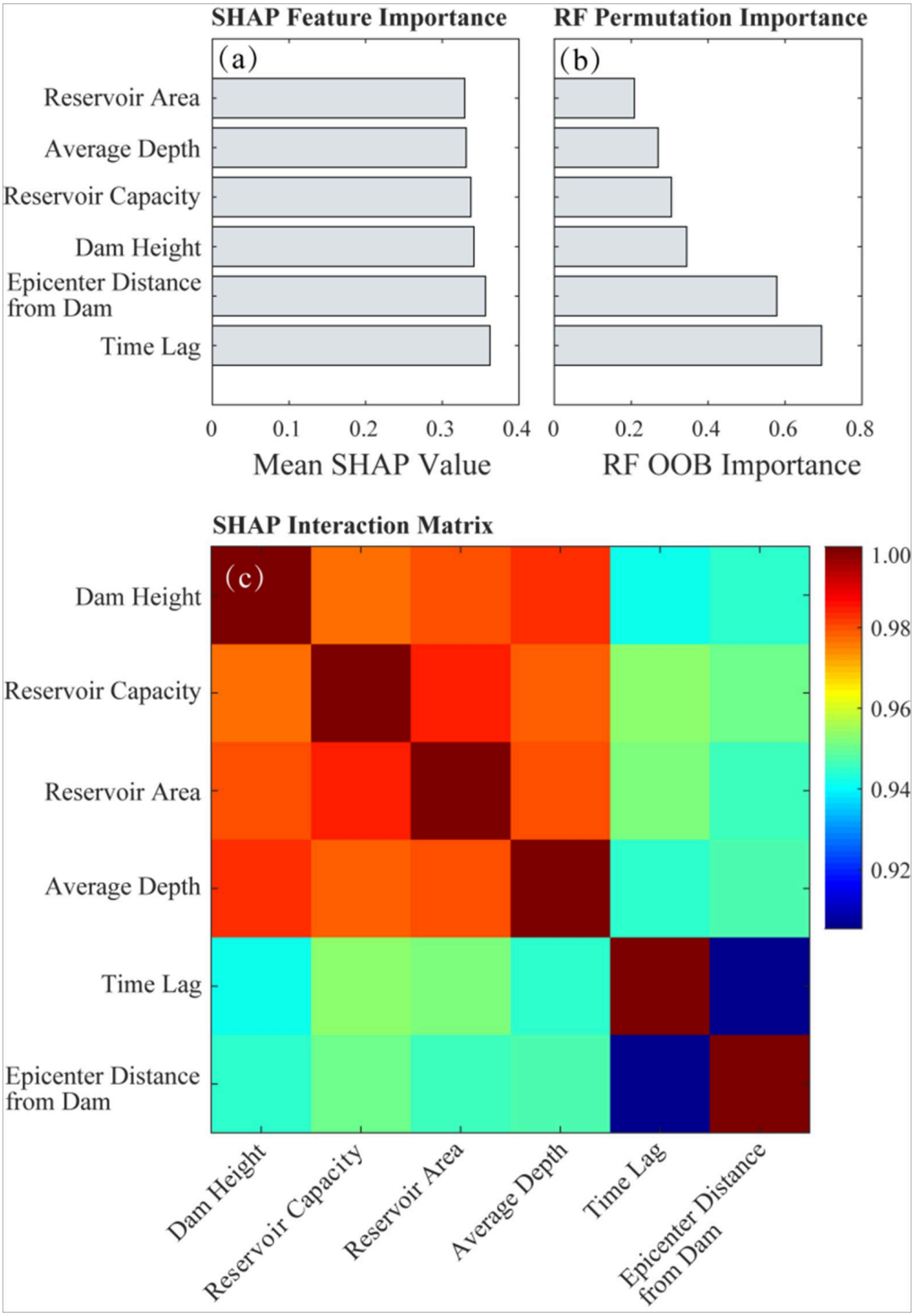


Figure 3

Analysis of the importance and independence of earthquake-inducing factors. (a) SHAP Feature Importance Analysis: Global importance ranking of features based on Shapley values. The height of the bar chart represents the mean absolute SHAP value for each factor. (b) Random Forest Intrinsic Importance: Feature importance calculated using the Out-of-Bag (OOB) permutation method. The values represent the increment in prediction error (standardized scale). (c) SHAP value correlation matrix of

earthquake-inducing factors: The color in the matrix indicates the Pearson correlation coefficient of SHAP values between different features.

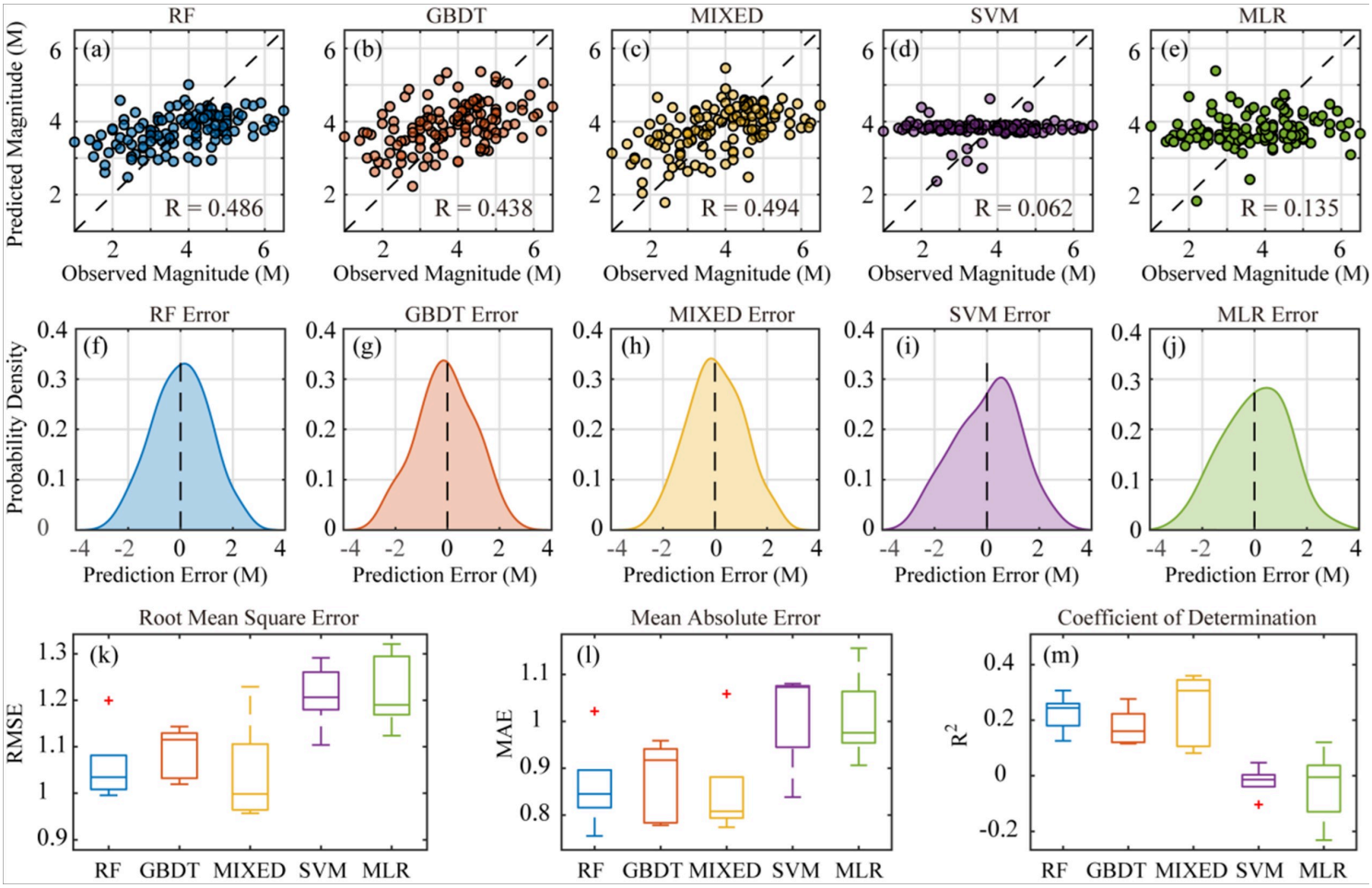


Figure 4

Comprehensive comparison of prediction performance of machine learning models for maximum magnitude (M_{max}) of reservoir-induced earthquakes. (a~e) The first row of subplots displays scatter plots of predicted versus observed values for the five models: Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Hybrid Model (MIXED), Support Vector Machine (SVM), and Multiple Linear Regression (MLR). The diagonal dashed line represents the ideal prediction line, and the Pearson correlation coefficient (r_p) is annotated. (f~j) The second row of subplots shows the probability density distributions of prediction errors for the corresponding five models. The dashed line marks the zero-error reference line.The third row presents boxplots of the statistical metrics: (k) Root Mean Square Error (*RMSE*), (l) Mean Absolute Error (*MAE*), (m) Coefficient of Determination (R^2). In the boxplots, the median line is shown, and the box represents the interquartile range (IQR, 25%–75%). The whiskers extend to the boundaries of 1.5×IQR, and red crosses indicate outlier points beyond the whiskers.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalMaterial.docx